A Comparison of Usability Evaluation Methods: Heuristic Evaluation versus End-User Think-Aloud Protocol – An Example from a Web-based Communication Tool for Nurse Scheduling

Po-Yin Yen, RN, MS[1], Suzanne Bakken, RN, DNSc[1,2]
[1]School of Nursing & [2]Dept of Biomedical Informatics, Columbia University, NY, NY

Abstract

We evaluated a web-based communication tool for nurse scheduling using two common usability evaluation methods, heuristic evaluation and end-user think aloud protocol. We found that heuristic evaluation performed by human-computer interaction (HCI) experts revealed more general interface design problems, while end-users' think-aloud protocols identified more obstacles to task performance. To provide the most effective and thorough evaluation results, a combination of heuristic evaluation and end-user think-aloud protocol is recommended.

Keywords:

Usability evaluation, heuristic evaluation, think-aloud protocol, staffing and scheduling, nursing

Introduction

Usability evaluation is a method for identifying specific problems with usability of products [1, 2]. The benefits of attending to usability issues through iterative evaluation include improved predictability of the products, greater productivity with fewer user errors, better match with user needs, and savings in development time and cost [1, 3-5]. A variety of usability evaluation methods have been used to detect usability problems related to technology, including heuristic evaluation, cognitive walkthrough, cognitive task analysis, think-aloud protocol, and usability surveys each offering a particular perspective. Participants range from HCI experts with no domain knowledge to those similar to actual users with domain knowledge to actual system users. In this study, two common usability evaluation methods, heuristic evaluation with HCI experts and think aloud protocol with actual system users, were compared.

Background

Heuristic evaluation

Heuristic evaluation was proposed by Nielsen as a usability inspection method and is guided by heuristic principles [6]. HCI experts discover system usability problems by detecting unmet heuristic principles, i.e., heuristic violations. As a discount usability engineering method, heuristic evaluation uses relatively few evaluators to detect the majority of usability problems [7, 8].

Think-aloud protocol

Think-aloud protocol was developed by Lewis in 1982 to understand cognitive process [9]. It encourages users to express out loud what they are looking at, thinking, doing, and feeling, as they perform tasks [9]. This allows observers to see and understand the cognitive processes associated with task completion. Using actual users or intended users as the participants in the think-aloud protocol provides a closer view of how users use the system and reveals practical usability problems related to task performance [10].

The purpose of the study is to compare the results from HCI heuristic evaluation and end-user think-aloud protocol.

Methods

A web-based communication tool for nurse scheduling, Bidshift, was assessed using two usability evaluation methods. The innovative tool is designed to address the nursing shortage by increasing the efficiency and effectiveness of the scheduling process. It allows nurse managers to announce open shifts throughout their organization and staff nurses to request shifts for which they are qualified based upon their profile. If more than one individual requests the same open shift, nurse managers are able to select a nurse based on her/his experience or working hours (e.g., not exceeding hospital overtime policy).

Heuristic evaluation and user think-aloud protocol were conducted to evaluate the usability of the web-based communication tool. Results were compared to understand the different perspectives offered by the techniques.

Heuristic evaluation

With no training on the system, HCI experts (n=5) were asked to perform four tasks for two Bidshift interfaces, Nurse Manager Interface (NMI) and Staff Nurse Interface (SNI): post/request a shift, award/check a requested shift, search for a nurse/shift, view report/shift schedule. They were also encouraged to explore other aspects of the system. Each expert completed a heuristic evaluation checklist based on Nielsen's principles [6] that included definitions, subquestions, and an overall 5-point Likert rating scale for each principle (from 0: no usability problem to 4: usability catastrophe). Expert's were encouraged to write comments to further explicate their rationale for their ratings.

End-user think-aloud protocol

Participants were recruited via personal contact from a single community hospital in the Philadelphia area. Inclusion criteria were: 1) nurse manager or staff member (Registered Nurse or Patient Care Technician), 2) experience using the BidShift Open Shift Management tool for at least 6 months on a monthly basis, and 3) hospital employee pre- and post-BidShift implementation and for > 1 year. Following informed consent, participants were asked to think aloud as they completed four sub-tasks associated with the open shift management process depending on their role of manager (M) or staff (S): 1) post/search for a shift, 2) award/request a shift, 3) search for a nurse/shift, and 4) view report/schedule. Their utterances and screenshots were recorded using Morae^{TM} software. After task completion, participants were asked three questions: 1) What do you like the most about the system and why? 2) What do you like the least about the system and why? 3) Do you have any suggestions for improving the system? Participants were probed to gather additional information regarding functionality, features, processes, user interface, user-system interactions and manager-staff communication. Data collection time ranged from 30 to 60 minutes. Data were managed and coded using Morae^{TM}. Participants' utterances were summarized and categorized into the ten heuristic principles for comparison with heuristic evaluation.

Results

HCI experts and system users agreed on some aspects of system usability, but exhibited a different focus for their usability concerns. HCI experts primarily commented on the general interface design such as the navigation in the system. On the other hand, system users commented on interface problems that impeded their task accomplishment. Table 1 shows the comments made from experts and end-users categorized by the heuristic principles.

Experts and system users both agreed on cosmetic and minor usability problems related to the heuristic principles of "User Control and freedom", "Consistency and Standards", "Error Prevention", "Flexibility and efficiency of use" and "Recognition rather than recall". Most users commented that it took a relatively small amount of time to learn how to use the system. Both groups of evaluators agreed that the system did not "Help Users Recognize, Diagnose, and Recover from Errors".

Experts identified several flaws in "Visibility of System Status" and "Aesthetic and Minimalist Design", however, only one user commented on the busy information in the system. Other users did not identify problems related to system visibility and aesthetic design. Also, experts commented more frequently on the interface features in "Match Between System and The Real World ". Nevertheless, several comments from users pointed out the mismatch between their expectations and system functions and how this mismatch led to inconvenience in task performance. Furthermore, experts rated the lack of a help page as a major usability problem. Experts looked for help page, which was not in place for the heuristic evaluation, when they encountered task performance difficulties. On the contrary, even though system users had access to a help page, they did not use it. Instead, they used a trial and error approach to figuring out how to do things.

System users also commented on system functions and how effective they perceived the function to be in achieving a particular task. Users were satisfied when a function was both effective and efficient for accomplishing a task [11].

Discussion

Experts versus end-users

Studies have demonstrated that both experts and end-users are effective in revealing usability problems, but that they capture different usability perspectives [12]. The findings of this are consistent with prior studies that suggest that HCI experts reveal more general interface problems while end-users identify severe interface obstacles to their task performance. Consequently, usability problems identified by experts but not by end-users are more likely to be interface features but less relevant to impact on task performance [13]. In other words, HCI experts disclose more of "ease of use" issues while end-users disclose more of "usefulness" issues.

HCI experts were not intended end-users of the system they evaluated, so they identified overall interface design problems. End-users described issues in priority to their impact on task performance. These findings are consistent with a study that compared four usability techniques and reported that even though experts found more problems than end-users, end-users were good at avoiding low-priority problems [14]. This can also be explained by the hypothesis that if a usability problem does not impact end-users' task completion, it may be less influential over time as users get used to the system. However, these usability problems do not disappear and usually frustrate new users [15].

Experts identified problems related to “Visibility of System Status” and “Aesthetic and Minimalist Design”, while users vividly described how lack of “Match between System and the Real World” task accomplishment. Such usability problems are unlikely to be discovered by heuristic evaluation [16].

Table 1. Heuristic Evaluation vs. End-user Think-aloud Protocol

1. Visibility of Status
Expert: Heuristic Evaluation Checklist rating score - NMI = 2.4; SNI = 1.4 *“The requesting shifts portion is not labeled well once one goes inside.”; “Not all functions have visual feedback in menus or dialog boxes about which choices are selectable”; “Awarding shifts and approving profile has an unusual system feedback. Wish it would explicitly say at top what task I am performing, not just ‘Main Page’”* **End-user:** (none)
2. Match between System and the Real World
Expert: Heuristic Evaluation Checklist rating score - NMI = 1.6; SNI = 1 *“Is there any specific order internally defined for the values in the list box, e.g., hospital, level of care? It doesn’t look like an alphabetical order.”; “it is not clear to me what happens when I click ’Inactivate’, are they gone for good or will I be able to retrieve the information later?”; “(Some actions) need some spontaneous instructions or guidance or definition, e.g. ‘use this page when…’, ‘shift points are …’”; “I expect a search function in the top right corner and quick search violates this convention.”* **End-user:** *“Our staff is still not used to looking at splitting into a partial shift. So they will pick it sometime thinking they are picking 3 to 7, but not looking at 3 to 11. So I got into some problems with that. When there is partial comes up. I don’t know what they are doing on their side of it, but they are not going further with that. [...] If it says 3 to 11, they are thinking they are working 3 to 7 and sometimes they just hit it and then they are surprised that they are not working 3 to 7. [...] they are not used to deal with partial shifts.”; “Because right now, I just stay on my unit, but in order to find out if you want to work on another unit. I didn’t notice that [...] I just push ‘Search’, I thought that you would go to everywhere. But I did not know that was my own unit. Then after a while, I realized that you have to push one of these to find another shift in different hospitals or different units in the hospital. That’s where only thing I think people will get confused at. They didn’t know that.”; “I’ve looked at this too to see if I can use it. [...]I don’t get to see. This to me is harder to look at and to move around than my schedule book. Only because I am probably more comfortable with this schedule book. That’s the way we have done for three years.”; “I think it is a little confusing that on the calendar, it does show you the time you are working. It doesn’t actually tell you like where you are, where you pick up. [...] I wish I would have like next to it ‘5C’(the location), something like that [...] because we usually just print out the calendar.”*
3. User Control and Freedom
Expert: Heuristic Evaluation Checklist rating score - NMI = 1.25; SNI = 1.6 *“User control and freedom is satisfactory”; “Does the system automatically log me out when I don’t use it for a certain period of time? Will all changes that I have made be saved?”; “Tabs make it easy.”* **End-user:** *“The first thing I do when I open Bidshift to look at is I notice what shifts are starting within 72 hours and I would want to pick those first”, “Usually when I come early in the morning, I check all my bid shifts. [...] I also pull shifts to see what my needs are to make sure they are on. I also double check the staff profile waiting to be approved.”; “I usually go in with a few days in mind, and times. [...]And usually, you could find the shift for what day you’re on. I look at my own unit first, and then if not I’ll look for other telemetry units”*
4. Consistency and Standards
Expert: Heuristic Evaluation Checklist rating score - NMI = 0.6; SNI = 0.4 *“Blue links appear to be too close in color spectrum to black”; “I don’t see big difference between ‘Quick Search’ in the main page tab and ‘Search’ by criteria in Search tab, except allowing multiple selections.”* **End-user:** (none)
5. Help Users Recognize, Diagnose, and Recover From Errors
Expert: Heuristic Evaluation Checklist rating score - NMI = 1.8; SNI = 1.4 *“I think this is a big issue. I did not know how to/what to type to recover from this error at post a shift”; “The error message was not much helpful. I never succeed in finding any shifts posted, although there are postings if you go and check out in a manager’s screen”; “Providing a help message for multiple selections (for search) is user friendly.”* **End-user:** *“See, it says ‘Sorry, but that shift id is invalid. Please try again’. But it doesn’t tell me which shift I am working, but I am in the schedule. It does that a lot after staffing puts it in. So I am like, I don’t know where I am working. It has it on another screen. I have to go to ‘MyShift’. It will say it over here [...] So it’s like I have to keep on clicking different link to see the information I want, instead of just being there.”*

6. Error Prevention
Expert: Heuristic Evaluation Checklist rating score - NMI = 1.4; SNI = 1.4 *“Sort of – retract my request. However, you can retract a request and there is no way to cancel or undo the retraction.”; “I find ‘Exit this screen’ confusing and inappropriate (return to prior screen). ‘Go Back’ and ‘Done’ are also confusing to me”* **End-user:** (none)
7. Recognition Rather Than Recall
Expert: Heuristic Evaluation Checklist rating score - NMI = 1.5; SNI = 0.8 *“Red shows shifts expiring, but when awarding shifts, color is not used well.”; “It really took me a moment to figure out how to find staff details.”* **End-user:** *“I don’t know what this mean (pointing at the calendar tabs).”*
8. Flexibility and Efficiency of Use
Expert: Heuristic Evaluation Checklist rating score - NMI = 1.25; SNI = 1.6 *“There is a search feature and the user can make the search more specific by selecting more search options.”; “No keyboard shortcut.”; “The system provide tab and shift-tab function.”* **End-user:** *“This is ‘Create a shift’, which is the same as ‘PowerPost’, just it can only do one shift at the time. You saw me how I do a bunch (through PowerPost), and we like a bunch” , “You can see now I have already awarded 111 shifts, and this schedule only has been out for two weeks [...] I have awarded more than half of what I need. [...] It exceeded my expectation.”*
9. Aesthetic and Minimalist Design
Expert: Heuristic Evaluation Checklist rating score - NMI = 1; SNI = 0.8 *“I found it overwhelming at first.”; “The unit calendar is well designed, but in my opinion, the other tabs are too busy.”; “I would try to reduce the view shift table, maybe delete one of the columns – it’s a lot of information.”; “The view just needed to be redesigned. The red text is slightly difficult to read.”; “Orange alert on the awarded shift page is difficult to read.”* *“’Search’ and ‘Quick Search’ options are confusing.”* **End-user:** *“It does give you a lot of access, so you have to be very careful [...] There is a lot of information on the screen. Maybe because I am lack of experience.”*
10. Help and Documentation
Expert: Heuristic Evaluation Checklist rating score - NMI = 2.75; SNI = 2.75 *“It took some time for me to figure out how to award shifts, could not figure out how to crate shifts. Help would be useful.”; “Did not see help page except FAQ list.”; “Should have a help page.”* **End-user:** *“I don’t think those features (Help) particularly useful in those system.”; “I never see that down there to use it (Help).”*

Note: NMI: Nursing Manager Interface; SNI: Staff Nurse Interface

Human-centered distributed information design

Human-centered distributed information design (HCDID), proposed by Zhang, recommends four levels of analysis to improve human-centered system design: user analysis, functional analysis, task analysis and representative analysis [17]. Not surprisingly, we found that HCI experts contributed predominantly to representative analysis. However, their contributions to task analysis and functional analysis were smaller as they were not the actual system users and were not as able to recognize the usability problems impeding task performance in the real world. On the other hand, users are good evaluating their own preferences and impediments to task performance, but are less able to articulate issues with the representation of the information in the system. For example, in this study they did not necessarily look for ‘Help’ provided by the system, but they discovered ways to interact with the system. The perceptions of HCI experts are useful to improve the interface to be more intuitive to new users, but problems related to system functions found under real-world conditions are unlikely to be discovered by heuristic evaluation. Hence, to provide the most effective and thorough usability evaluation result, combination of usability evaluation techniques from both expert and system user perspectives is recommended [10, 18].

Study limitations

This study has several limitations. First, due to the small sample size, the study findings may be limited in generalizability. Second, unlike to HCI experts, system users were not explicitly instructed to look for problems related to heuristic principles. , Users may have provided more interface design problems if heuristic principles were provided as a stimulus. However, our think aloud protocol approach provided freedom to end-users to express their concerns. Providing heuristic principles may restrict users from exploring real world usability problems. Lastly, the comparison of the two usability evaluation methods was limited to a single system, the web-based communication tool for nurse scheduling. The comparisons of expert and system user perspectives may vary based upon the type of system evaluated.

Conclusion

Two common usability evaluation methods, heuristic evaluation and end-user think aloud protocol, were compared. Results found that heuristic evaluation performed by HCI experts revealed more general problems, while think-aloud protocol of system users identified more usability obstacles to task performance. To provide the most effective and thorough usability evaluation results, a combination of usability evaluation techniques from both expert and system user perspectives is recommended.

Acknowledgement

Thanks to Douglas Hughes and Catherine DiNardo from Main Line Hospital who facilitated the system user data collection and to Concerro (http://www.concerro.com/) for providing access to Bidshift for the heuristic evaluation. The study was supported by the Center for Evidence-based Practice in the Underserved (P30NR010677).

References

1. Dumas JS, Redish JC. *A Practical Guide to Usability testing*. Revised ed. Portland: Intellect Ltd; 1999.

2. Yao P, Gorman PN. Discount usability engineering applied to an interface for Web-based medical knowledge resources. *Proceedings / AMIA ... Annual Symposium. AMIA Symposium.* 2000:928-932.

3. Nielson J. *Usability Engineering.* Cambridge: Academic Press; 1993.

4. Wiklund ME. *Usability in Practice*. Cambridge: Academic Press; 1994.

5. Rubin J. *Handbook of usability testing*. New York: John Wiley & Sons; 1993.

6. Nielsen J. Ten Usability Heuristics. 2005; http://www.useit.com/papers/heuristic/heuristic_list.html. Accessed March 4, 2008.

7. Desurvire H, Kondziela J, Atwood ME. What is gained and lost when using methods other than empirical testing. *SIGCHI Bulletin.* 1991;23(4):125-126.

8. Desurvire H, Lawrence D, Atwood M. Empiricism versus judgement: comparing user interface evaluation methods on a new telephone-based interface. *SIGCHI Bulletin.* 1991;23(4):58-59.

9. Lewis C. *Using the "think aloud" method in cognitive interface design.* New York: IBM 1982.

10. Holzinger A. Usability engineering methods for software developers. *Communications of the Acm.* Jan 2005;48(1):71-74.

11. Yen P, Bakken S. Usability Testing of a Web-based Tool for Managing Open Shifts on Nursing Units. Paper presented at: The 10th International Congress on Nursing Informatics2009; Helsinki, Finland.

12. Lai T-Y, Larson EL, Rockoff ML, Bakken S. User acceptance of HIV TIDES--Tailored Interventions for Management of Depressive Symptoms in persons living with HIV/AIDS. *Journal of the American Medical Informatics Association.* Mar-Apr 2008;15(2):217-226.

13. Doubleday A, Ryne M, Springett M, Sutcliffe A. A comparison of usability techniques for evaluation design. *Proceedings of the 2nd conference on Designing interactive systems* 1997:101-110.

14. Jeffries R, Miller J, R., Wharton C, Uyeda K. User interface evaluation in the real world: a comparison of four techniques. Paper presented at: Proceedings of CHI; April 27 - May 02, 1991; New Orleans, Louisiana.

15. Kjeldskov J, Skov MB, Stage J. A longitudinal study of usability in health care - does time heal? *Studies in Health Technology & Informatics.* 2007;130:181-191.

16. Jeffries R, Desurvire H. Usability testing vs. heuristic evaluation: was there a contest? *SIGCHI Bulletin.* 1992;24(4):39-41.

17. Zhang JJ, Patel VL, Johnson KA, Smith JW, Malin J. Designing human-centered distributed information systems. *Ieee Intelligent Systems.* Sep-Oct 2002;17(5):42-47.

18. Virzi RA, Sorce JF, Herbert LB. A comparison of three usability evaluation methods: heuristic, think-aloud, and performance testing. *Proceeding of the Human Factors and Ergonomics Society 37th Annual Meeting.* 1993:309-313.